group meeting 2022-10-25

Parsimonious representation of biological twitching trajectories.

Intro: twitching motility

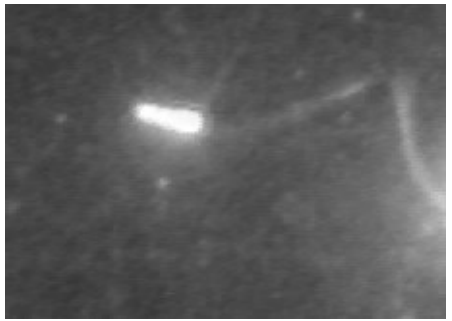


Fig 1. Skerker & Berg, 2001

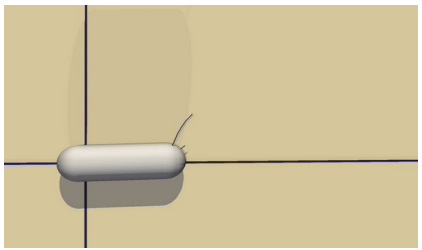


Fig 2. Simulation

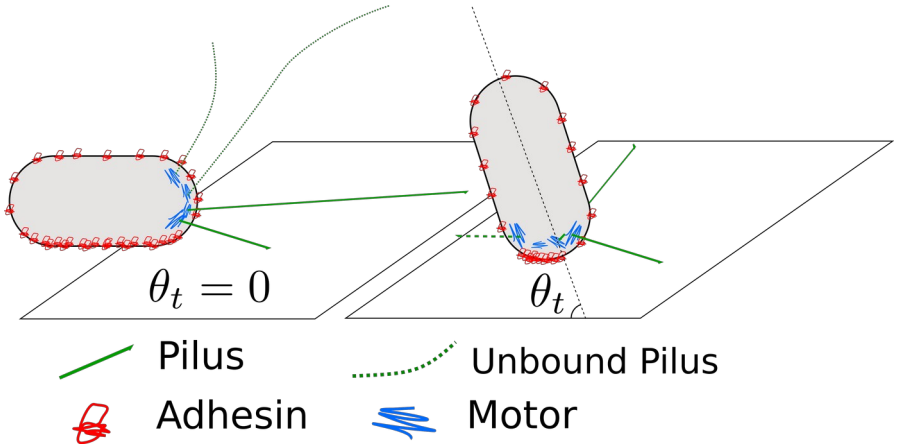


Fig 3. Simplified Drawing

Tracking data from twitching experiments

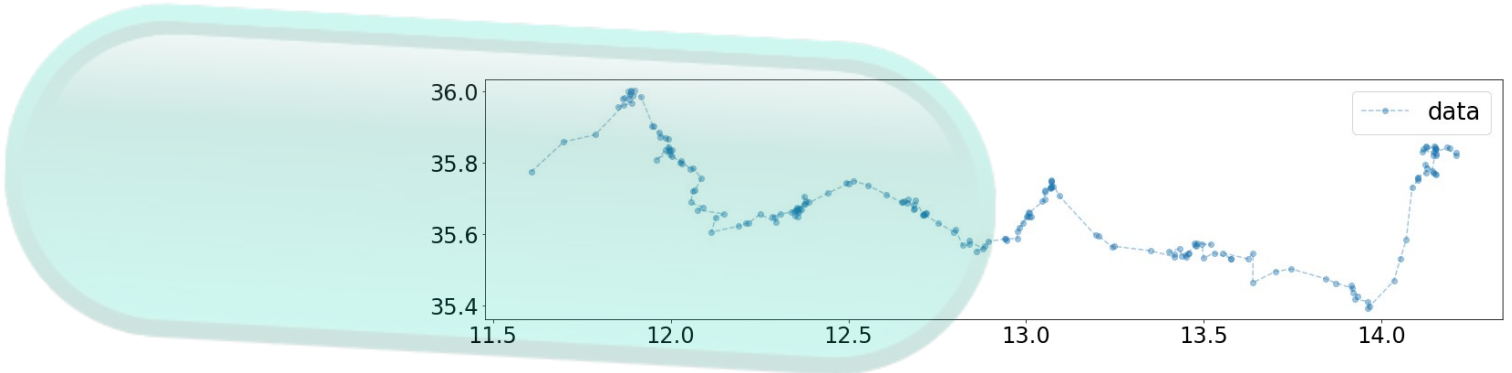


Fig 1. 20 seconds of Tracking Data (μm)

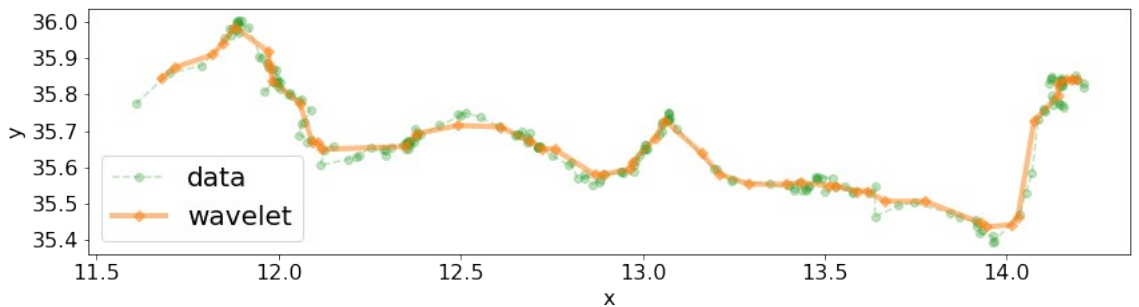


Fig 2. tracking data with wavelet smoothing

Question: Can we get more useful, more interpret-able information from this trajectory data?

Piecewise Linear Solve

consider N measurements taken at equal time intervals.

$$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$$

and define a *partition* of this data as a series of M indices

$$(k_1, \ldots, k_M),\ \mathrm{k}_1 < k_2 < \ldots < k_M$$
$$k_1 = 0, k_M = N$$

for any sub-sequence of the data we could do a linear fit

$$\boldsymbol{y} = m\boldsymbol{u} + \boldsymbol{c}$$

where distance to the line of a single data point is

$$d_i = ||\boldsymbol{c} + \boldsymbol{u} \cdot (\boldsymbol{x}_i - \boldsymbol{c})\boldsymbol{u} - \boldsymbol{x}_i||$$

and its common to obtain **u**, **c** by least squares optimisation. Minimise

$$\phi_2(k_1, k_2) = \sum_{i=k_1}^{k_2} d_i^2$$

Another (probably worse) option is to minimise

$$\phi_\infty(k_1, k_2) = \max\left(d_{k_1}, d_{k_1+1}, \ldots, d_{k_2}\right)$$

Lets define a global cost function

$$\Phi(k_1, ..., k_M) = \sum_{k_i} \phi(k_i, k_{i+1})$$

Draw your attention to the following two issues.

Issue 1: This cost function decreases with increasing M, in fact

$$M \to N, \Phi \to 0$$

In regression and machine learning, this is called “overfitting”. We want M to be just large enough to capture the piecewise linear features of the data and no larger.

Issue 2: Global optimisation by searching all possible {M, k_1, …, k_M} becomes quickly infeasible for even moderate data size N.

Greedy Recursion

A efficient method will help us build up intuition

$$\text{set } M = N,\ k_1 = 0, k_2 = 1, \ldots, k_M = N,$$

So that each of the N-1 linear fits contains only two points, for which a perfect fit is always available.

$$\phi_\infty(i, i+1) = 0$$

Now consider joining any one of the data points to its neighbours, we could write this operation in pseudocode as

$$k_{i+1} \leftarrow k_{i+2}$$
$$M \leftarrow M - 1$$

there are N-2 possible join operations on the first iteration. At each interation we choose to join the pair with the minimum cost

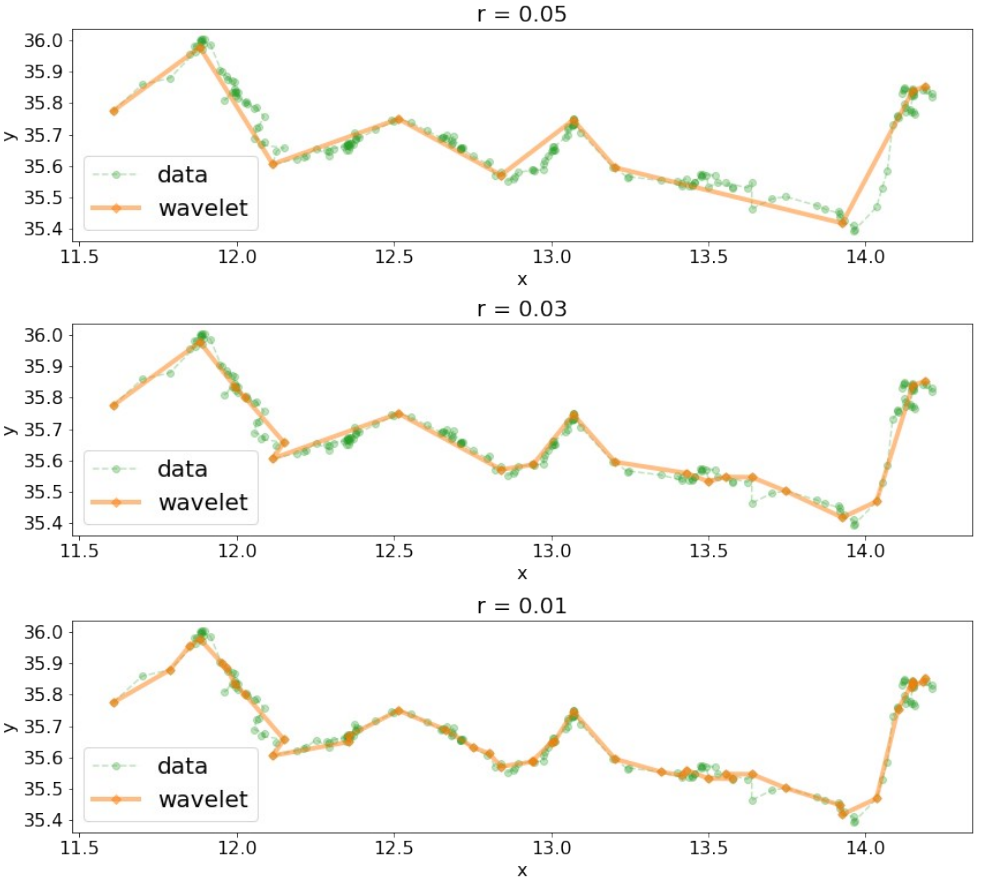
$$\phi_\infty(k_i, k_{i+2})$$

Lets choose a threshold cost and use it as a stopping criteria (otherwise we will continue to join data until M = 1).

$$\text{Stop if } \phi_\infty(k_i, k_{i+2}) > r$$

Finally obtain a connected piecewise linear model from the partition by simply selecting breakpoints from the data itself.

$$\text{breakpoints } \{\boldsymbol{x}_{k_i}\}$$



Greedy Recursion on Simulated data

This method forms the basis for more advanced algorithms.

It also works well on simulated twitching trajectories which have no noise but still have some minor non-linearities.

In this case were interested in comparing simulated and real twitching trajectories, so we output data using the experimental timestep of 0.1 seconds.

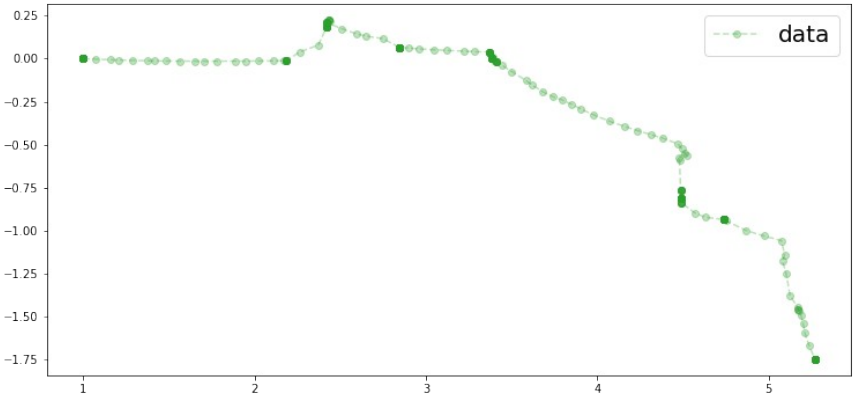
We assumed the measurement errors in real data follow a normal distribution and estimated sigma using a heuristic method.

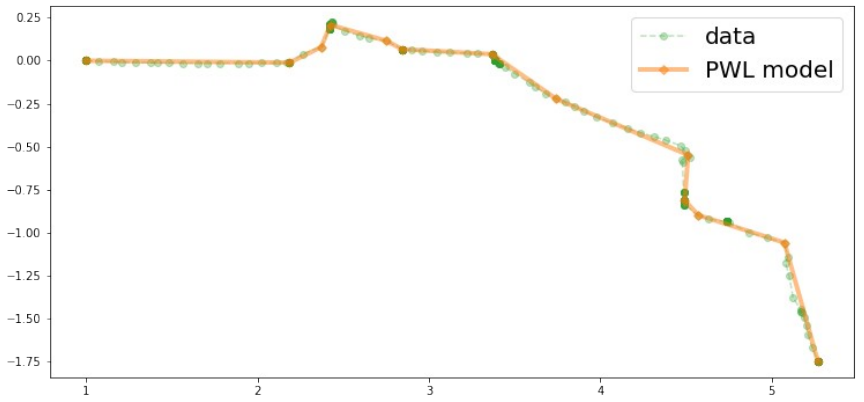
$$\mathcal{N}(0, \sigma^2),\ \sigma = 0.012$$

Roughly, if we want to use the ϕ_∞ cost function to extract linear features of the data with a high confidence we could use the inverse cumulative distibution function to set r, e.g.

$$r = CDF^{-1}_{\mathcal{N}_\sigma}(0.99) = 0.028$$

Which says that there is a 1% chance that a measurement error will be greater than 0.028.
(2% If we consider that the error could be on either tail)





r = 0.030

Simulated Annealing

A more expensive approach is to stochastically (but exhaustively) search the space of partitions

$$\{M, k_1, ..., k_M\}, M \leq N$$

for the minimal partition that nonetheless satisfies a cost condition for every piecewise element.
That is

$$\text{Minimise } M \text{ such that } \phi(k_i, k_{i+1}) < r, \forall i \quad *$$

Lets define the following three random transitions.

$$\text{Move: } k_i \leftarrow k_i \pm 1, \text{ maintain } k_{i-1} < k_i < k_{i+1}$$

$$\text{Join(i): } k_j \leftarrow k_{j+1}, \forall (j > i), M \leftarrow M - 1$$

$$\text{Split: opposite of Join}$$

We can always accept random transitions which decrease M while maintaining the constraint.

Its possible that we will sometimes want to increase M, or break the constraint so long as we eventually converge a partition that satisfies the constraint.

Simulated annealing is a Monte Carlo optimisation algorithm which uses a decreasing “temperature” T to control the probability that an unfavourable transition is accepted.

Transitions are accepted if

$$P(S, S_{\text{new}}, T) \geq \text{Uniform}(0, 1)$$

Where S is the state of the partition, and P would satisfy

$$P \geq 1 \text{ if } M_{\text{new}} < M \text{ and } *$$

$$0 \leq P < 1 \text{ if } M_{\text{new}} \geq M \text{ or } !*$$

$$P \to 0 \text{ as } T \to 0 \text{ if } M_{\text{new}} \geq M \text{ or } !*$$

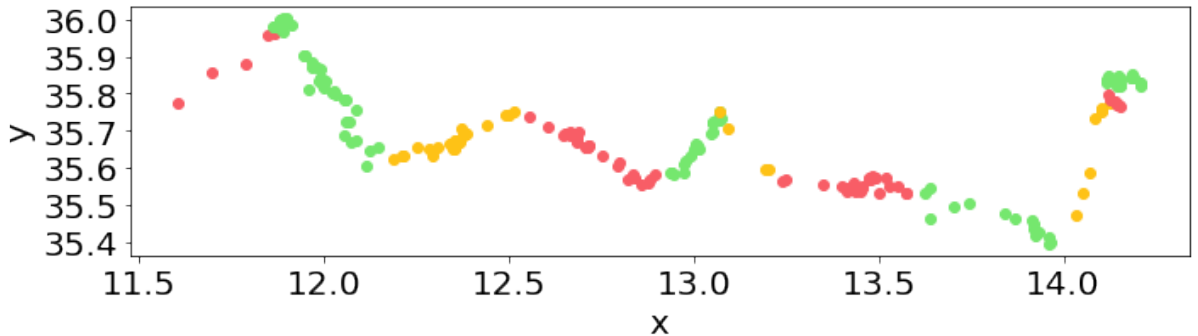
The temperature is then gradually reduced to 0 during the optimisation.

Simulated Annealing (Visual Aid)

< Libreoffice couldn’t handle this GIF >

Connected PWL model

- Solving for an optimal partition is not quite the same as solving for the piecewise linear model itself.

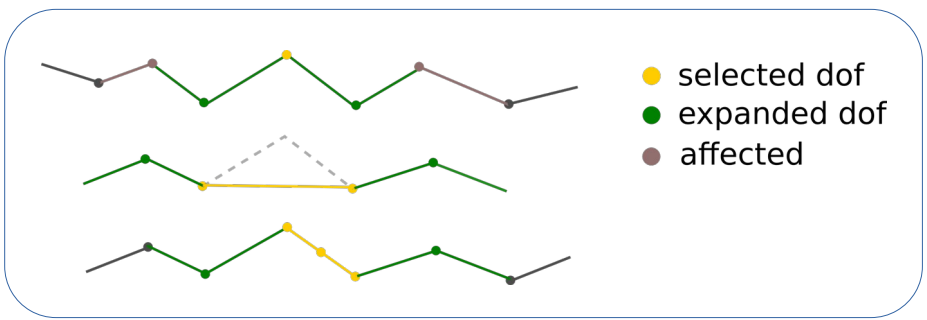


- But the alternative, optimising a connected piecewise linear model, is quite expensive.

random move

local solve

Random moves



Minimum Description Length

We can solve for a connected piecewise model by reinterpreting the threshold r and defining a new global cost function using an idea from information theory.

$$\text{let } \Phi_{\mathrm{DL}} = M + \sum_i c_r(d_i),\ c_r(d_i) = 1 \text{ if } d_i > r \text{ else } 0$$

The second part is the number of “outliers”, points which are distance r or further from the candidate piecewise curve.

This “description length” encodes information about both the goodness of fit and the number of piecewise components. We can search for a global minimum.

Unfortunately the description length is an integer valued function making it difficult to optimise. I use a trick here:

$$\text{Given } \{M, k_1, \ldots, k_M\}, \text{ minimise } \Phi_2, \text{ then compute } \Phi_{\mathrm{DL}}$$

And other trick is to minimise the tuple

$$(\Phi_{\mathrm{DL}}, M)$$

Which means that removing piecewise segments is preferred if doing so leaves the description length unchanged.

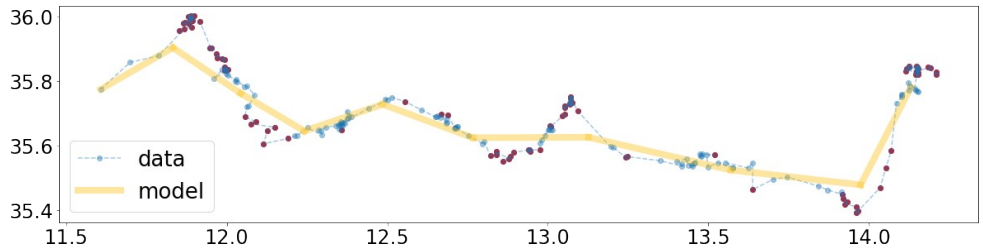


Fig 1. Initial guess with outliers (red) and points close to the model (blue)

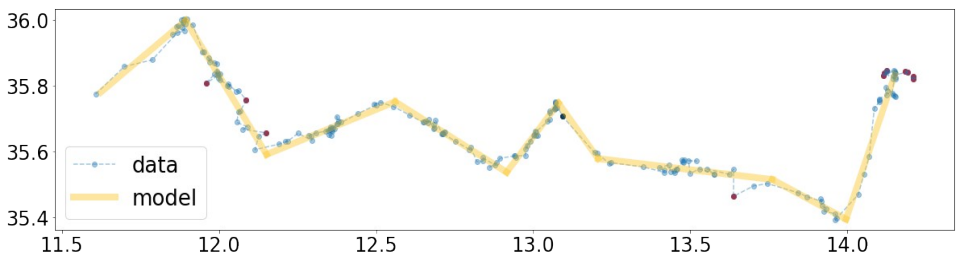


Fig 2. Minimum description length model.

Minimimum Description Length

Pro

- Insenstive to outliers
- Conceptually simple

Con

- Arbitrary weighting of outliers and model segments
- Introduces a discrete threshold into otherwise continuous variables.
- Difficult to optimise directly
- Threshold throws away some information

Pro/Con

- The r threshold is quite robust, but is it robust enough?

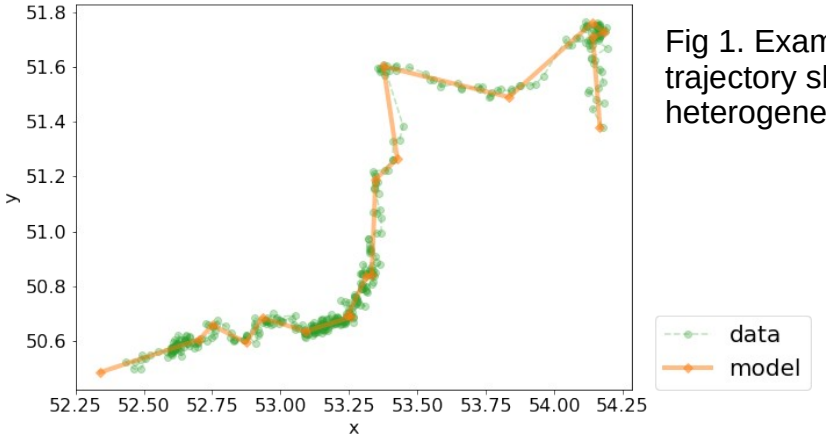
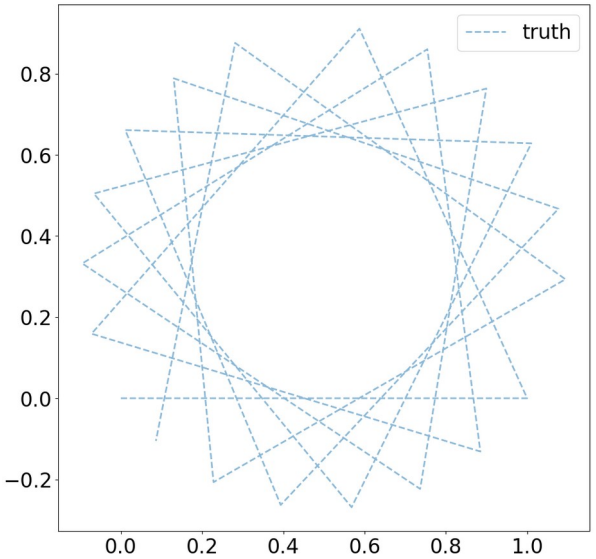
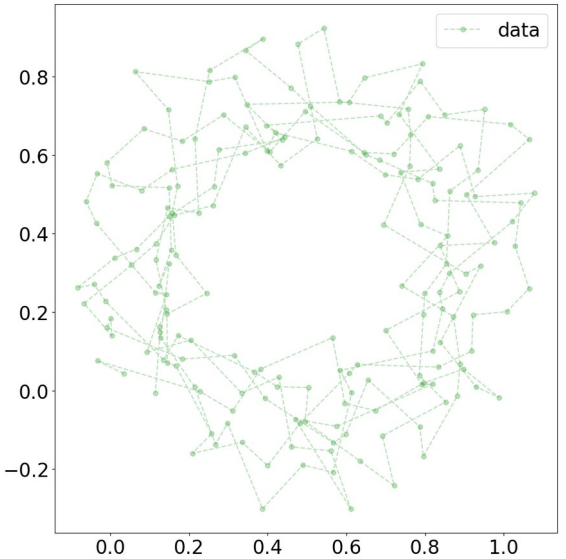
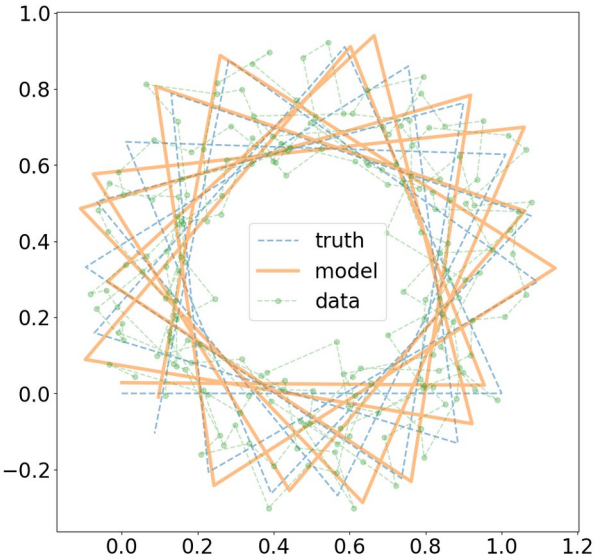


Fig 1. Example trajectory showing heterogeneous velocity.

Synthetic Data Example







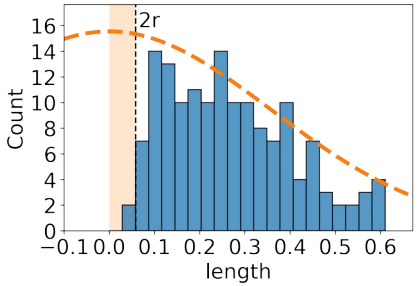
- Generate 10 random points per line segment.

sigma = 0.1
step length = 1.0
p_threshold = 0.98
r = 0.23
sampling_dx = 0.1

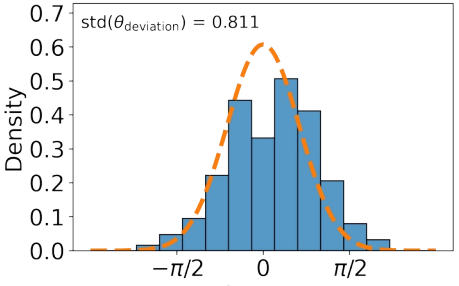
Statistics

This experimental trajectory has ~2000 data points

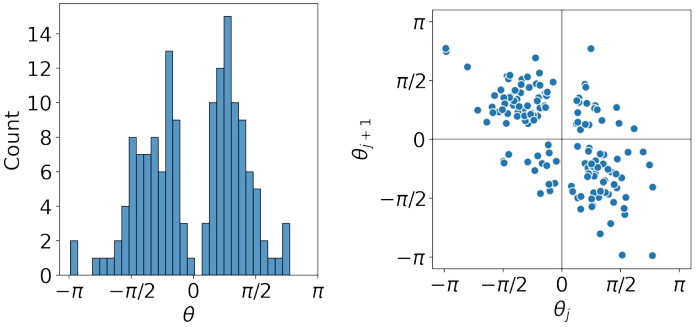
- Segment length distribution



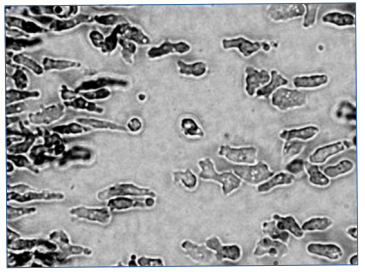
- Angle between body axis and velocity



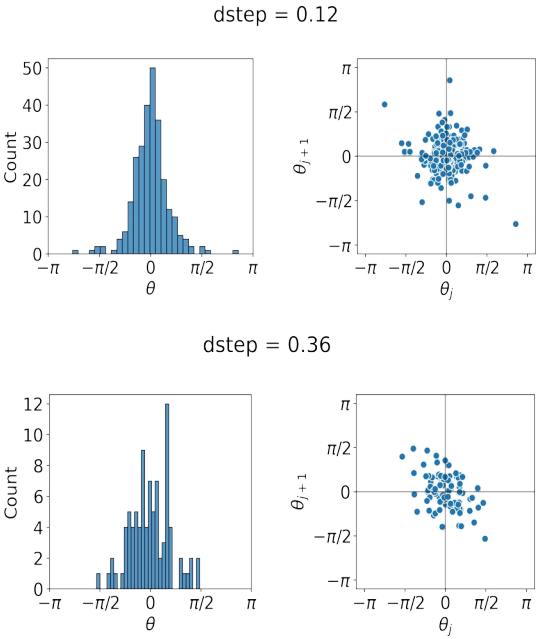
- Sequential linear displacements are anticorrelated (corrcoef = -0.65)



- This behaviour is shared with dictyostelium

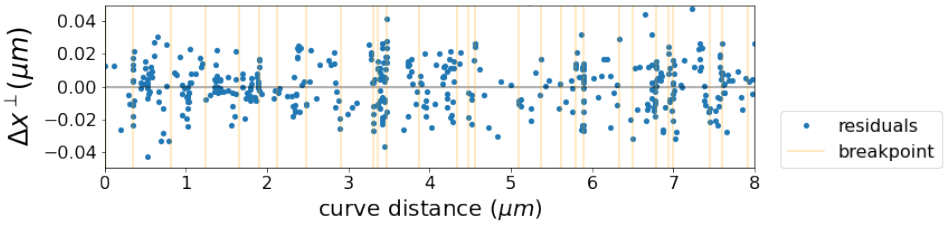


- The effect can be seen by a simple coarse graining with length 0.36 μm

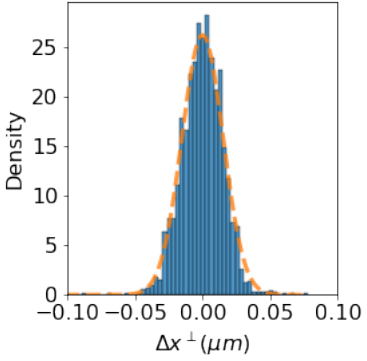


Residual Analysis

- Residuals vs. curve distance
- measurements clustering around breakpoints indicates low velocity/pausing between linear segments



- normal distribution of residuals



Estimated std : 0.0126

Residual std : 0.0152

“unexplained” deviation

0.0152 – 0.0126 = 0.0026

- Short timescale non-linearities would show up in the correlation between residuals.

$$\rho(\Delta x_i, \Delta x_{i+1}) = 0.285$$

$$\rho(\Delta x_i, \Delta x_{i+2}) = 0.087$$

- Durbin Watson statistic. Values less than 2 indicate a positive correlation.

$$d = \frac{\sum_i (\Delta x_i - \Delta x_{i-1})^2}{\sum_i \Delta x_i^2} = 1.43$$

- Very short timescale non linearities implies that the bacteria makes displacements which are beyond our spatial and temporal resolution to detect.

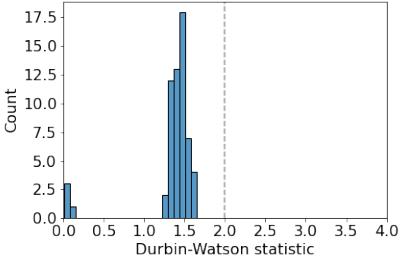
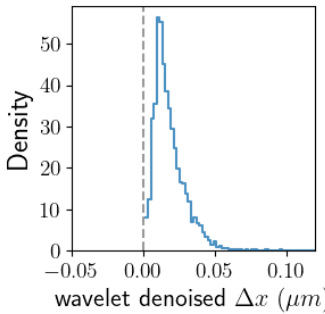
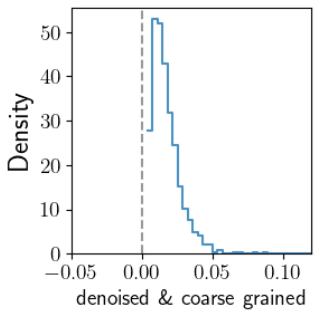


Fig. Population statistics for similar trajectories (N=60)

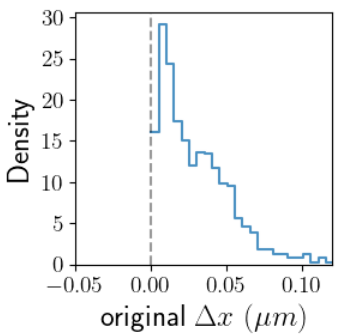
Velocity Distributions

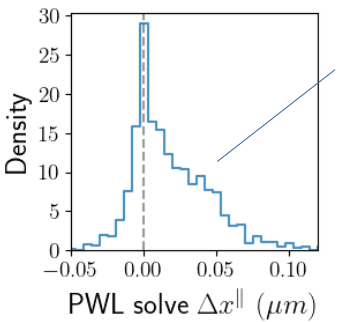
- Fanjin et al. report a bimodal velocity distribution for particular twitching trajectories.
- For comparing simulated and experimental data, I usually use a denoised (wavelet transform) and coarse grained displacement distribution





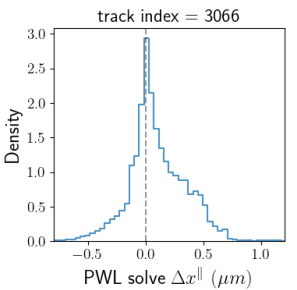
- but biomodal velocity distribution is easier to spot in unprocessed 0.1s distplacements.

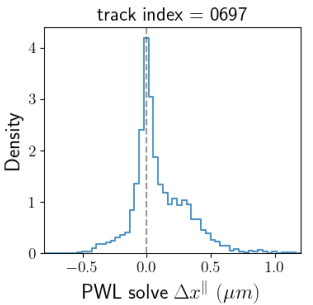


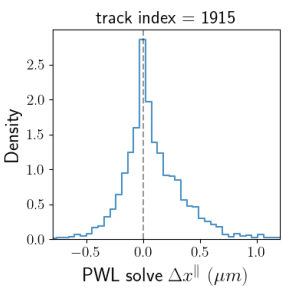


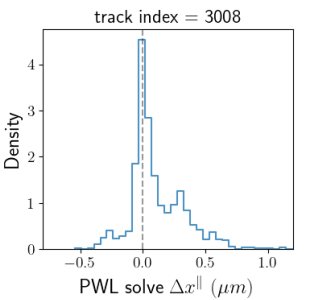
obtained by mapping data onto PWL solution

- Bimodal velocity distribution is a generic property? Plotting distributions for similar trajectories...



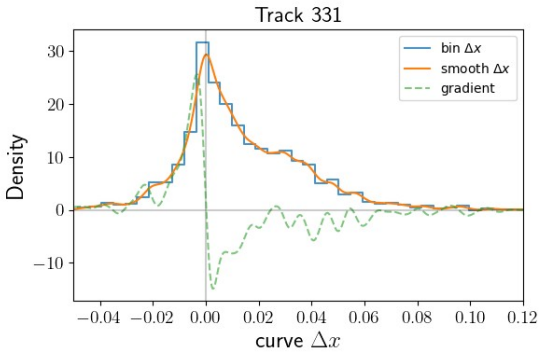






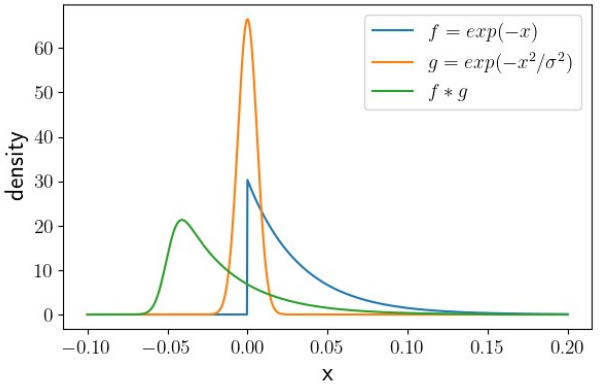
Bimodal velocity: generic property of twitching or not?

Idea 1: peak fitting



- Look for peaks in smoothed distribution or its gradient.
- Continue by post processing the results for significant peaks.

Idea 2: model fitting



- Construct “mixture models” with e.g. exponential part and a gaussian part.
- Convolve models with the error distribution and then compare to to smoothed data.

That’s Enough